Exploring Inference Techniques in Difference-in-Differences Analysis

Sneha Roy , Biswajit Palit

University of Bonn

February 23, 2024

Abstract

A growing literature on inference in Difference-in-differences design has been pessimistic about hypothesis tests of correct size especially with fewer groups. In this paper, we propose and evaluate several methods through Monte Carlo simulations, which are capable of providing correctly sized tests even with a small number of groups. Moreover in this paper, we evaluate the efficacy of these methodologies in detecting genuine effects, which presents a more serious challenge in DD estimation than obtaining a correct test size. Through extensive Monte Carlo simulations, we establish the small sample properties of the estimators and evaluate their performance across a variety of data-generating processes. Furthermore, we apply the findings from our simulations to a real-world dataset, demonstrating the empirical utility of the best-performing estimator we identify.

Keywords: Difference-in-differences, small number of groups, power, small sample properties

1 Introduction

Differences-in-Differences (DD) estimation has gained popularity as a method for estimating causal relationships. DD involves identifying a distinct intervention or treatment, often the implementation of a new law. This approach compares the difference in outcomes after and before the intervention for groups affected by the intervention to the same difference for unaffected groups. DD estimates and their standard errors are most often derived from using Ordinary Least Squares (OLS) in repeated cross-sections (or a panel) of data on individuals in treatment and control groups for several years before and after a specific intervention. Formally,

$$y_{igt} = \gamma_0 + \gamma_1 \alpha_g + \gamma_2 \delta_t + c \cdot w_{igt} + \beta T_{gt} + \epsilon_{igt} \tag{1}$$

where α_g represents the state fixed effects and is equal to 1 if the state is treated and 0 if the state is not treated. δ_t represents the time fixed effect, which is equal to 1 in the post-treatment period and 0 in the pre-treatment period. The interaction of the state and the time effects gives us the treatment intervention variable T_{gt}, which is equal to 1 only in the states which are treated in the post-treatment period and 0 otherwise. w_{igt} represents the relevant individual controls and ϵ_{igt} is the error term. The estimated impact of the treatment is then given by the OLS estimate of $\hat{\beta}$.

This specification represents a common generalization of the most basic 2X2 DD setup, which typically involves two periods and two groups. Intuitively, this DD setup compares four cell-level means, where only one cell is treated. The difference in the expected outcome for the control group (the group that is not treated) between post and pre-treatment periods illustrates the time trend, while the difference in the expected outcome of the treatment group minus the time trend serves as the treatment effect estimate. However, it is important to note that this setup is valid only under the very restrictive assumption that changes in the outcome variable over time would have been the same in both treatment and control groups in the absence of the intervention. This assumption is commonly referred to as the parallel trends assumption. It is important to note that DD estimation does not rely on assumption of homogeneous treatment effects. When treatment effects are homogeneous, DD estimation yields average treatment effect on the treated (ATT). When impacts change over time (within treated units), DD estimate of treatment effect may depend on choice of evaluation window.

In this paper, we argue that the estimation of equation (1) is subject to a potentially severe serial correlation problem in practice. To assess the extent of this issue, we investigate how DD performs on placebo laws, where treated states and the year of passage are chosen randomly. As these laws are fictitious, a significant ”effect” at the 5 percent level should be observed roughly 5 percent of the time. However, we find significantly higher rejection rates of the null hypothesis of no effect.

Consequently, we proceed to suggest alternative methods of estimating the average treatment effect. We examine whether these methods yield correct test sizes under various robustness checks, including a reduction in the number of groups and misspecification of the error structure.

Initially, we analyze the placebo laws using a real-world dataset from the Current Population Survey (CPS) of the US, where no actual laws have been enacted. Subsequently, we utilize synthetically generated data with a homogeneous AR1 structure to analyse the effect of the placebo laws in a more general setup. In the final section of the paper, we conduct a small empirical analysis on a real-world dataset to assess the impact of WTO ascension on the agricultural imports of developing countries.

2 Overrejection in DD estimation

2.1 Data :

We obtained microdata on women in the 50 states of the US from the fourth interview month of the Merged Outgoing Rotation Group (MORG) of the Current Population Survey (CPS) for the years 1980 to 2000, inclusive (21 years). The sample comprises nearly 900,000 observations. Our focus is on women of working age, those between 25 to 50 years old. The extracted data includes information on weekly earnings, education level, age, and state of residence. For our analysis, we have considered the log of weekly earnings as the dependent variable. Among the 900,000 women in the original sample, approximately 443,000 report strictly positive weekly earnings. This results in 1,050 state-year cells (50 states multiplied by 21 years), with each cell containing an average of a little more than 400 women with strictly positive earnings. At the individual level, we include age and education dummies[1] as covariates.

We use a two-step procedure to compute $\hat{\beta}_{OLS}$ (the estimated treatment effect) from micro-data regression using equation (1).

Step 1 : In the first step we aggregate the data from microdata into 1050 state-year cells. We conduct a regression using the microdata where the outcome variable y_{igt} is regressed on individual-level control variables w_{igt}. This regression captures the relationship between the outcome and individual-level covariates such as age and education levels. We then obtain the residuals from this regression, representing the part of the dependent variable that is not explained by the individual-level covariates. Essentially, these residuals now capture the variation in the dependent variable attributable to state and time effects, as well as any treatment effects that may have occurred. We calculate the mean of these residuals across states and time to obtain the state-year cells containing the aggregated values of the residuals (represented by $\hat{Y}_{gt}$).

Step 2 : In the second step, we estimate the equation

$$\hat{Y}_{igt} = \alpha_g + \delta_t + \beta T_{gt} + \epsilon_{igt} \tag{2}$$

[1] The education dummies we use can be categorized by : Up to Grade 10, High School, Master's Degree, and Doctorate Degree

where T_{gt} represents the treatment variable. This equation models the relationship between the aggregated residuals $\hat{Y}_{gt}$ and the treatment variable while accounting for state and time effects (α_g and δ_t respectively) as well as any unexplained variation ϵ_{gt}. The OLS estimate $\hat{\beta}$ represents the estimated treatment effect.

2.2 Problem of Serial Autocorrelation :

The correlogram analysis of wage residuals has highlighted a notable issue of serial autocorrelation within the dataset. We estimate the first, second, and third autocorrelation coefficients for the residuals from a regression of $\hat{Y}_{gt}$ on the state and year dummies. The autocorrelation coefficients are then obtained by a simple OLS regression of the residuals on their corresponding lagged residuals. The calculated first-order autocorrelation coefficient stands at 0.347, exhibiting high significance up to the 1 percent level. Additionally, the third autocorrelation coefficient registers at 0.106, maintaining significance at the 1 percent level. These findings collectively suggest a notable presence of serial autocorrelation within the dataset.

2.3 Placebo Intervention :

To quantify the problem induced by serial autocorrelation in the DD context we randomly generate placebo laws that affect some states and not others, in a staggered manner. We introduce staggered placebo laws to selected states (the treatment group). The commencement of treatment is chosen randomly from the years 1985 to 1995 to ensure an adequate number of observations before and after the treatment.

In the first step, we choose at random exactly half the states (25) and designate them as treatment states. Then for each designated treatment state, we pick a year from a uniform distribution between 1985 to 1995 and designate them as the treatment starting year for that state.

The intervention variable T_{gt} is then defined as a dummy variable which equals 1 for the treated states after the treatment starting period and 0 otherwise. [2]

We then estimate DD equation 1 using OLS on these placebo laws. This generates an estimate of the law's effect and a corresponding standard error. To understand how well conventional DD performs, we repeat this exercise 5000 times, each time drawing new laws at random. If the conventional DD provided an appropriate estimate for the standard error, we should expect that the null hypothesis of no effect i.e, β = 0 should be rejected exactly 5 percent of the times, when we use a threshold of 1.96 for the absolute t-statistic. The above exercise gives us an idea about the Type I error.

A small variant of this exercise allows us to assess the power of the test. In this section, we have introduced a 0.05 log point, i.e, we have added 0.05 to y_{igt} only for the treated state time cells. [3] The idea is to simulate a 5 percent effect on the data and then we have proceeded with the data aggregation as before. We want to calculate the power of the power of the test by calculating the number of times the null hypothesis of no effect (β= 0) gets rejected against the alternate hypothesis of ($\beta \neq 0$) out of 500 simulations. By repeating this exercise, we can assess how often DD finds an effect when there is one.

From the first row of table 1 in the appendix, we report that the null hypothesis of no effect is rejected 29.64% out of 5000 simulations, which deviates from the expected 5 percent significantly. One reason for this gross over-rejection is that OLS fails to account for the correlation between state–year cells. In other words, OLS assumes that the variance-covariance matrix for the error term is diagonal. However, in reality, it might be block diagonal, with correlations

[2] Ensuring parallel trends by construction : To ensure that there is no structural bias between the treatment and control states, we have randomized the assignment of placebo treatments to the states. Randomization helps us achieve causal relationships by ensuring that the assignment of treatments is based on chance rather than any pre-existing characteristics or biases. So, we can attribute any observed outcome differences to the treatment only. Thus we completely do away with the possible endogeneity of the intervention themselves.

[3] We introduce the effect on the micro level data and thereafter perform the aggregation.

within each state-year cell, but no correlation across states. That is, for each state the value of the residual of one year is correlated with the value of residual from the previous years as we observed from the wage correlogram. We observe an average standard error value of 0.009 using OLS while the standard error of $\hat{\beta}$ distribution comes to be 0.017.

This suggests that the presence of autocorrelation in the errors introduces a dramatic downward bias in the computed standard errors. To illustrate this problem, we construct a different type of intervention variable. As before, we again randomly choose exactly half of the states to be the treatment group. However, we randomly select only 10 dates to be the intervention dates for each simulation. The treatment variable is now defined as 1 if the observation relates to a treated state and only if it belongs to one of the 10 intervention dates, 0 otherwise. Therefore, the treatment variable is repeatedly turned on and off with the value yesterday giving us no information about the value today. Thus, the serial autocorrelation in the treatment is essentially eliminated. From the second row, in Table 1, we see that the rejection rate of OLS has significantly reduced to 0.0528 percent and the standard error of the $\hat{\beta}$ distribution becomes 0.009. This suggests that the bias in the standard errors is indeed induced by the presence of serial autocorrelation of the error terms.

3 Data Generation for Simulation study :

To understand how the over-rejection vary with the degree of serial autocorrelation in the dependent variable we experiment with various autocorrelation parameters in a synthetically manufactured data. We employ a homogenous AR1 data generation process according to the following specification:

$$Y_{gt} = \alpha_g + \delta_t + \rho \cdot Y_{gt-1} + \epsilon_{gt} \tag{3}$$

where α_g and δ_t are the state and time fixed effects, randomly drawn from $\mathcal{N}(0,1)$ distribution. ρ is the autocorrelation parameter. Y_{gt-1} is the lagged outcome and ϵ_{gt} denote the error term which is randomly chosen from an $\mathcal{N}(0,1)$ distribution.

We observe from Table 2 in the Appendix that the rejection rate increases with the increase in the autocorrelation parameter. Interestingly, we see a correctly sized test when the autocorrelation parameter is 0. But even at moderate levels, that is when ρ is 0.2 or 0.4 the rejection rates are almost two to four times as large as they should be. We also observe an increase in average standard errors along with an increase in rho, which leads to lesser precision of the treatment coefficient estimates and hence a flatter beta distribution with the increase in the autocorrelation parameter as seen by an increase in the standard error of $\hat{\beta}$ distribution in Table 2.

4 Alternative Estimation Methods

In this section, we evaluate the performance of different alternative estimation methods, that have been proposed in literature to deal with the problem of serial autocorrelation. To do so, we evaluate the impact of the placebo treatment using the CPS Aggregate data and the homogenous AR1 manufactured data. We also analyse the power of these methods to detect a 5 percent real effect. Also, we vary the number of states to understand how these methods perform in the case of a smaller number of groups.

4.1 Cluster Robust Standard Errors :

Our starting point for alternative estimation methods is based on the variants of Liang and Zeger's (1986) cluster robust standard error (CRSE) estimator. The formula for a cluster-robust variance matrix is as follows:

$$\hat{\sigma}^2_{CR} = (X'X)^{-1}\left(\sum_{g=1}^{G} X_g u_g u_g' X_g'\right)(X'X)^{-1} \tag{4}$$

where X represents the matrix of independent variables (year dummies, state dummies, and treatment dummy). X_g is the regressor matrix for group g. And u_g is the vector of regression residuals for group g.

This estimator is consistent for fixed panel length and the Wald statistics based on it are asymptotically normal as g tends to infinity (Kezdi 2002). The traditional CRSE estimator uses an $\mathcal{N}(0,1)$ distribution to compute the critical values. This is especially a problem when the number of states is small and the distribution of the treatment coefficient follows a t distribution with g-1 degrees of freedom. The t distribution assigns a higher mass at its tails which increases the absolute critical values. Thus, to fall in the rejection region with the t distribution, the t statistic has to be higher making it a more conservative test as compared to the traditional CRSE estimation method. Table 3 of the Appendix summarises the result of CRSE method taking its critical values from $\mathcal{N}(0,1)$ distribution and Table 4 summarises the results of CRSE $t(g-1)$. We can observe that with a lesser number of states, the $\mathcal{N}(0,1)$ critical values yield considerably higher rejection rates than the $t(g-1)$ critical values.

An intriguing observation arises when comparing the average standard error of the Cluster-Robust Standard Error (CRSE) methods with traditional OLS method. It is noted that the average standard error values for both CPS data and manufactured data simulations exhibit a substantial increase with the CRSE approach. This increase in standard error leads to a reduction in rejection rates. Additionally, the CRSE $t(g-1)$ method can detect a 5 percent effect 81 percent of the times with the CPS dataset along with providing a correctly sized test. Despite having fewer groups, the CRSE $t(g-1)$ method consistently yields appropriately sized tests. However, it is observed that the Type 1 error increases when employing the CRSE $\mathcal{N}(0,1)$ approach.

In the synthetically generated data, a notable observation emerges: there is a substantially lower power to detect real effects compared to the CPS data. Specifically, we find that an effect as sizable as 5 percent becomes obscured by the inherent noise present in the data. This phenomenon is further underscored by the observation of higher average standard errors and standard errors within the distribution of $\hat{\beta}$.

4.2 Residual Aggregation :

Another way to address the problem of serial autocorrelation is to ignore the time series aspect when calculating standard errors.First, we regress the outcome variable on state and year dummies and get the residuals. Then, we split these residuals into a two-period panel for each state: one for years before the treatment and one for years after. We calculate the average residual for each state in each group to obtain one value in each panel for each state. In the post-treatment group, we indicate the treatment dummy as 1, and in the pre-treatment group, as 0. Finally, we run a simple OLS regression on this two-period panel.

For 50 states, this method provides us with a correctly sized test as seen from Table 5.The downside of this procedure is that it performs poorly when the number of states decreases, the rejection rates for 10 states, and 6 states are 8.6% and 10.9% respectively, which are considerably higher than an expected 5%. Although for 50 states, this method yields a marginally higher power than the CRSE methods.

4.3 Wild Cluster Bootstrapping :

With a few groups, an alternative to relying on asymptotic normality is that we can recover the distribution of the test statistic via bootstrap. Referring to Cameron, Gelbach, and Miller (2008), we employ the wild cluster bootstrapping procedure to obtain correctly sized tests. In our bootstrap implementation, we depart from the conventional approach of resampling individual observations. Instead, we resample clusters of residuals, each cluster containing all observations within it. These resampled clusters are then adjusted by a constant drawn from a two-point distribution, specifically, 1 and -1, each with an equal probability of 0.5, known as the Rademacher distribution. This bootstrap method gives us a correctly sized test even for a small number of states. Although this method gives us good results, both in terms of the test size and power of the test, even for lesser number of states as seen from Table 6 of Appendix, this method is extremely computationally expensive.[4]

4.4 Parametric Approach - Feasible Generalized Least Squares (FGLS) :

Our final approach to address the serial correlation problem is to use used Feasible Generalised Least Squares Method. A natural way to proceed is to assume a homogenous AR k structure for the residuals. In this paper, we have assumed a homogenous AR 2 process for the same. The first step involves regressing the dependent variable on the independent variables and computing the residuals, then regressing these residuals on up to 2 lags to compute the autocorrelation parameters. These parameters are then used to perform a GLS transformation of both the dependent and the independent variables. The final regression involves a standard OLS regression of the transformed variables to estimate the treatment coefficient.

In this respect, two issues arise. First, estimates of the parameters obtained by regressing the residuals on k lags (2 in our case) can be inconsistent with T fixed, due to the presence of fixed group effects. Hansen (2007) derives a bias correction which is consistent as G tends to infinity and develops the asymptotic properties of the FGLS estimator. However, this correction may not be robust to a small number of states. Second, since this process assumes an underlying

[4] For each simulation, we are resampling the residuals 400 times.

homogenous AR k process of residuals, one may be worried about the misspecification of the error structure.

In our analysis, we utilized both Feasible Generalized Least Squares (FGLS) and Bias Corrected FGLS methodologies. Table 7 in the appendix presents the results obtained from FGLS, while Table 8 outlines the outcomes of the bias-corrected version. We observe that both of these approaches consistently yield correct test sizes when applied to CPS data, even as the number of states decreases. However, when considering homogenous AR1 data, we note that the type 1 error rate tends to increasecrease with a reduction in the number of states for FGLS. In contrast, BC-FGLS maintains the correct test size across varying numbers of states, even for homogenous data.

FGLS demonstrates superior performance compared to all previously discussed methods in detecting a genuine effect of 5 percent, especially when the number of states is sufficiently large (achieving 84.8% power with 50 states). Additionally, the bias-corrected version also performs well recording marginally higher power than the CRSE $t(g-1)$ method.

5 Robustness Analysis

In this part of our analysis, we conduct a Monte Carlo simulation study using two other data-generating processes where we misspecify the error process. First, we use a heterogenous AR 1 process of generating synthetic data following the equation 3, where, for each state the autocorrelation coefficient rho is drawn from a uniform distribution $\mathcal{U}(0.2, 0.9)$. This breaks down the homogenous autocorrelation structure in the outcome variable across states. We also use a homogenous MA 1 process to generate the data following the equation:

$$Y_{gt} = \alpha_g + \delta_t + \theta \cdot \epsilon_{gt-1} + \epsilon_{gt} \tag{5}$$

where the error of the current year depends on the error of the previous year for each state. Here θ represents the shock parameter and we assume homogeneity for all the states. For our analysis, we have considered theta to be 0.5. We have then provided the treatment variable just as discussed above.

From Table 9 in the appendix we observe that CRSE $\mathcal{N}(0,1)$ has given a slightly higher test size for homogenous MA1 data. Interestingly Residual aggregation fails the robustness analysis to provide the right test sizes and overestimates the type 1 error to be 0.092 for heterogenous AR1 process and 0.088 for homogenous MA1 process. Similarly, the FGLS method also fails to provide correct test sizes for both heterogeneous AR1 and homogeneous MA1 processes. However, the bias-corrected version remains robust to error mis-specification. Unsurprisingly OLS gives us the highest power among all the methods, however, its gross over-rejection of the null hypothesis is the reason why it is unsuitable for DD designs with serial autocorrelation. Except residual aggregation most of the methods provide correctly sized tests in the face of mis-specified error processes.

6 Empirical Analysis

Since its establishment in 1995, the WTO has seen 36 states or customs territories join as Article XII members, a process governed by negotiations outlined in the Marrakesh Agreement. These negotiations, conducted under the ”single undertaking”, require prospective members to adhere to all WTO agreements. Accession typically involves significant domestic reforms, with governments citing reasons such as economic restructuring and market transition. Economic arguments supporting WTO membership highlight its role in reducing trade barriers, promoting international trade, and fostering economic growth. The WTO’s Agreement on Agriculture, negotiated during the Uruguay Round, aimed to liberalize agricultural trade by reducing tariffs and subsidies. Additionally, the WTO provides a platform for resolving trade disputes related to agriculture and facilitates trade through measures such as customs streamlining.

This section examines the hypothesis that WTO accession positively impacted agricultural imports for Article XII members. We employ a difference-in-difference analysis, considering the staggered accession periods of these countries — for instance, China in 2001 and Ukraine in 2008.

6.1 Data :

To study the effects of recent WTO accession on total agricultural imports[5], we put together a large country-level panel dataset, which covers the period 1992-2015. As mentioned before, the 36 Article XII countries that ascended to the WTO post-1995 become our treatment group. We used the WTO developing country members that were already part of the GATT[6] to select the control group. [7] [8]

Subramanian and Wei (2007) discovered that while the GATT/WTO has generally benefited global trade, its impact on developing nations appears to be less pronounced. This suggests that membership in these organizations might not significantly alter the import behaviors of developing countries. Their findings echo Rose's (2004) assessment, indicating that developing nations have undertaken only modest trade liberalization efforts. Furthermore, Subramanian and Wei pointed out a trend where countries joining the GATT after the Uruguay Round tend to demonstrate greater openness in trade policies compared to long-standing developing country members. Consequently, to ensure a suitable comparison, we opted to use pre-Uruguay Round

[5] Sourced from Food and Agriculture Organisation database (FAOSTAT)

[6] The organisation was called GATT (General Agreement of Tariffs and Trade) before the inception of WTO in 1995

[7] In theory, the control group should include non-WTO members whose characteristics are similar to the acceding governments. For statistical reasons, this is not practical, because the group of countries outside the WTO is relatively small (35) and too heterogeneous to provide for a control group.

[8] We do not take into account the developed countries as part of our analysis because we want the control group to closely match the treatment group (the Article XII members are all developing nations)

developing country members of GATT (which later came to be known as the WTO after the Uruguay Round in 1995) as our control group for analysis.

6.2 The Model :

In the spirit of Leamer (1983), we have used a naive model for our analysis.

$$y_{igt} = \gamma_0 + \gamma_1 \cdot \alpha_g + \gamma_2 \cdot \delta_t + \beta T_{gt} + u_{igt} \tag{6}$$

The specification of the model remains same as before, where α_g represents the state effect and δ_t represents the time effect. The estimated impact of WTO ascension on the normalised value of Agri Imports is given by $\hat{\beta}$. We have performed this regression for each separate clusters of countries to understand the heterogenous treatment effect across the different clusters.

The clustering procedure looks at disaggregating the sample of WTO developing country members (both GATT and Article XII members) into a series of more homogeneous subsamples which may, for structural reasons, have different reactions to (i) economic shocks and trends; and, (ii) WTO accession treatment. The procedure, based on a series of structural socioeconomic and macroeconomic variables is implemented through two steps: (1) exclusion of outliers; and, (2) identification of clusters.[9]

The model is likely to suffer from the presence of autocorrelation in the residuals, and thus we cannot rely on the standard OLS inference. Thus we use the Bias-corrected FGLS method, since we see from our simulations, that it gives us a correctly sized test with a high degree of power, even with small number of states. This approach is particularly well-suited for our study, as we are conducting individual regressions for each cluster, thereby necessitating a robust method capable of handling autocorrelation for a small number of countries effectively.

[9] The Appendix subsection **??** provides detailed information on the procedure and its results.

6.3 Result of Data Analysis :

Our regression analysis looks at the impact of the ascension of WTO on Article XII members' total agricultural trade imports. As we see from **??** of the Appendix, for clusters 1 and 6 we find a positive and significant (at the 1% level) impact of WTO ascencion on normalized Agri Imports. For the remaining clusters we observe there is no significant result of the same. these results confirm the heterogenity of responses, and that the fact that a thorough investigation would require looking more at the economic characteristics of the acceding governments. Figure **??** of Appendix shows the heterogenity of responses to ascension across clusters.

Here we acknowledge, some limitations to our analysis I.e a relatively short time period to observe the impact of ascension since most of the countries acceded in the 2000s and 2010s. Also, we have considered only 3 years prior to the Uruguay Round in 1995, giving us a rather short time period to observe the pre-ascension outcomes and to ensure parallel trends. Moreover, this short time period was also included the 2008-2009 crisis, which might have affected the agricultural imports in ways which cannot be captured by the time effects only. Moreover Clusters 2, 4 and 5 and may suffer from treatment group imbalance[10] leading to a drop in power of FGLS to detect real effect.

7 Conclusion

Existing research has emphasized the challenges associated with obtaining accurately sized hypothesis tests, particularly when dealing with a limited number of groups. However, our findings suggest that this may not be the primary obstacle. Through Monte Carlo simulations, we aim to elucidate several key points. Firstly, we demonstrate the feasibility of deriving tests of

[10] the ratio of the size of treatment group to control group is very low

the correct size, even when confronted with a small number of groups. Importantly, we show that in many settings, achieving this is achievable using methods that are straightforward to implement. Secondly, we argue that the principal challenge in difference-in-differences designs with grouped errors lies in the inadequate power to detect genuine effects, especially when dealing with a fewer number of groups. Our analysis reveals that both the Bias-Corrected Feasible Generalized Least Squares (BC-FGLS) and the Cluster-Robust Standard Error $t(G-1)$ method yield correctly sized tests, even for a lower number of groups and in the presence of mis-specified error processes. Notably, as the number of groups decreases, the power of both the BC-FGLS and CRSE $t(G-1)$ methods diminishes sharply. While the wild cluster bootstrap also yields satisfactory results, it is computationally expensive.

8 References

1. Marianne Bertrand & Esther Duflo & Sendhil Mullainathan, 2002. ”How Much Should We Trust Differences-in-Differences Estimates?,” NBER Working Papers 8841, National Bureau of Economic Research, Inc.

2. Cameron, A. C., J. G. Gelbach, and D. L. Miller. 2008. “Bootstrap-Based Improvements for Inference with Clustered Errors.” The Review of Economics and Statistics 90 (3): 414–427.

3. Hansen, C. 2007. “Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects.” Journal of Econometrics 140 (2): 670–694.

4. Solon, G. 1984. “Estimating Autocorrelations in Fixed Effects Models.” National Bureau of Economic Research Technical Working Paper 32.

5. ANGRIST, J. D. AND J.-S. PISCHKE (2009): Mostly harmless econometrics: An empiricist’s companion, Princeton university press.

6. Kezdi, Gabor, ”Robust Standard Error Estimation in Fixed-Effects Panel Models” Working Paper, University of Michigan, 2002

7. Escaith, Hubert and Chemutai, Vicky, An Empirical Assessment of the Economic Effects of WTO Accession and Its Commitments (February 6, 2017). World Trade Organization, Economic Research and Statistics Division: Staff Working Paper ERSD-2017-05

8. Leamer, E. (1983) ‘Let’s Take the Con Out of Econometrics’ +The American Economic Review Vol. 73, No. 1 pp. 31-43

9. Rose. A (2004) ’Do we really know that WTO increases trade? American Economic Review’ 94 (1): 98-114

10. Subramanian A and S.J. Wei (2007) ’The WTO promotes trade strongly- but unevenly.’ Journal of International Economics 72(1):151-175

9 Appendix

9.1 Small Sample Properties of Various Methods

Table 1: : OLS assuming iid for CPS Data

Data	Rejection Rate		Average Standard Error	RMSE	Standard Error of $\hat{\beta}$	Bias
	No effect	5 percent effect				
CPS Data	0.2964	0.976	0.009	0.0014	0.017	0.0001
CPS Data : Serially Uncorrelated Laws	0.0528	0.998	0.009	0.0002	0.009	0.0001

Notes*

1. The data from the Current Population survey has been used. The data is for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1980 to 2000 inclusive. The data has been aggregated as outlined in section 2.1 of the paper. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.

2. All simulations have been performed for the significance level of 5%

3. The number of simulations for both rows is 5000, except for the 5% effect, for which it is 500. This reduction is due to the inclusion of the effect in the micro data, resulting in longer execution times.

4. The standard error of the rejection rates are calculated using the number of simulations using the formula $se(\hat{r}) = \sqrt{\hat{r}(1-\hat{r})/sim - 1}$. We are not explicitly reporting it, but this represents the accuracy of the rejection rate.

5. All regressions also include, in addition to the intervention variable, state and year fixed effects.

6. The average standard error is calculated by taking the mean of the standard error associated with the intervention variable for all the simulations.

7. The standard error of $\hat{\beta}$ is the standard error associated with the distribution of coefficient of the intervention variable obtained from the simulations.

8. The average standard error, RMSE, standard error of $\hat{\beta}$, and bias are reported for the simulations, calculating the rejection rate of no effect.

Table 2: : OLS assuming iid for homogeneous AR(1) data

Rho	Rejection Rate		Average Standard Error	RMSE	Standard Error of $\hat{\beta}$	Bias
	No effect	5 percent effect				
0	0.0492	0.053	0.117	0.166	0.117	-0.0001
0.2	0.096	0.099	0.118	0.194	0.14	-0.0003
0.4	0.1664	0.1734	0.124	0.255	0.177	-0.0003
0.6	0.272	0.287	0.141	0.246	0.251	0.0006
0.8	0.4204	0.4378	0.205	0.156	0.501	0.006

Notes*

1. The data has been generated following a homogenous AR(1) structure. The autocorrelation parameter used for the data generation is given in the column Rho.
2. All simulations have been performed for the significance level of 5%
3. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.
4. The number of simulations for each cell is 5000.
5. All regressions also include, in addition to the intervention variable, state and year fixed effects.
6. The average standard error is calculated by taking the mean of the standard error associated with the intervention variable for all the simulations.
7. The standard error of $\hat{\beta}$ is the standard error associated with the distribution of coefficient of the intervention variable obtained from the simulations.
8. The average standard error, RMSE, standard error of $\hat{\beta}$, and bias are reported for the simulations, calculating the rejection rate of no effect.

Table 3: : CRSE $\mathcal{N}(0,1)$ for CPS data and homogenous AR(1) data

N	Rejection Rate		Average Standard Error	RMSE	Standard Error of $\hat{\beta}$	Bias
	No effect	5 percent effect				
			A. CPS DATA			
50	0.05	0.802	0.017	0.001	0.017	0.0001
20	0.0578	0.414	0.027	0.031	0.027	0.0005
10	0.079	0.252	0.038	0.102	0.038	0.0003
6	0.11	0.236	0.048	0.031	0.05	0.0007
			B. HOMOGENOUS AR (1) DATA (ρ = 0.8)			
50	0.05	0.051	0.515	0.156	0.0501	0.006
20	0.059	0.061	0.82	0.355	0.818	0.021
10	0.0728	0.0716	1.58	0.64	1.163	-0.003
6	0.106	0.1082	1.47	0.161	1.543	0.006

Notes*

1. In panel A, data from the Current Population survey has been used. The data is for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1980 to 2000 inclusive. The data has been aggregated as outlined in section 2.1 of the paper. No effect represents the type 1 error and 5% effect represents the power of the test. For the type 1 error (No effect) the data is aggregated, hence, the number of simulations is 5000, and for the 5% effect (power) the number of simulations is 500. This is because the log point is implemented at the micro level of the data making it computationally very expensive and time-consuming to run a higher number of simulations.
2. All simulations have been performed for the significance level of 5%
3. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.
4. For panel B, the data has been generated using a homogenous AR(1) structure. This analysis uses only the autocorrelation parameter, rho 0.8. In this panel, all the simulations are run 5000 times because the data is generated in an aggregated form.
5. For panel A, we randomly choose the required number of states for each simulation without replacement. For panel B we just generate the data for the required number of states in the first place.
6. Here the critical values are computed from the $\mathcal{N}(0,1)$ distribution.
7. All regressions also include state and year fixed effects in addition to the intervention variable.
8. The average standard error is calculated by taking the mean of the standard error associated with the intervention variable for all the simulations.
9. The standard error of $\hat{\beta}$ is the standard error associated with the distribution of coefficient of the intervention variable obtained from the simulations.
10. The average standard error, RMSE, standard error of $\hat{\beta}$, and bias are reported for the simulations, calculating the rejection rate of no effect.

Table 4: : CRSE t(g-1) for CPS data and homogenous AR(1) data

N	Rejection Rate		Average Standard Error	RMSE	Standard Error of $\hat{\beta}$	Bias
	No effect	5 percent effect				
			A. CPS DATA			
50	0.0452	0.81	0.017	0.001	0.017	0.0001
20	0.0462	0.38	0.027	0.031	0.027	0.0005
10	0.0456	0.162	0.038	0.102	0.038	0.0003
6	0.055	0.148	0.048	0.031	0.05	0.0007
			B. HOMOGENOUS AR (1) DATA (ρ = 0.8)			
50	0.044	0.043	0.515	0.156	0.0501	0.006
20	0.046	0.047	0.82	0.355	0.818	0.021
10	0.0412	0.042	1.58	0.643	1.163	-0.003
6	0.049	0.0508	1.47	10.161	1.543	0.006

Notes*

1. In panel A, data from the Current Population survey has been used. The data is for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1980 to 2000 inclusive. The data has been aggregated as outlined in section 2.1 of the paper. No effect represents the type 1 error and 5% effect represents the power of the test. For the type 1 error (No effect) the data is aggregated, hence, the number of simulations is 5000, and for the 5% effect (power) the number of simulations is 500. This is because the log point is implemented at the micro level of the data making it computationally very expensive and time-consuming to run a higher number of simulations.
2. All simulations have been performed for the significance level of 5%
3. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.
4. For panel B, the data has been generated using a homogenous AR(1) structure. This analysis uses only the autocorrelation parameter, rho 0.8. In this panel, all the simulations are run 5000 times because the data is generated in an aggregated form.
5. For panel A, we randomly choose the required number of states for each simulation without replacement. For panel B we just generate the data for the required number of states in the first place.
6. Here the critical values are computed from the $t(g-1)$ distribution.
7. All regressions also include state and year fixed effects in addition to the intervention variable.
8. The average standard error is calculated by taking the mean of the standard error associated with the intervention variable for all the simulations.
9. The standard error of $\hat{\beta}$ is the standard error associated with the distribution of coefficient of the intervention variable obtained from the simulations.
10. The average standard error, RMSE, standard error of $\hat{\beta}$, and bias are reported for the simulations, calculating the rejection rate of no effect.

Table 5: : Residual Aggregation for CPS data and homogenous AR(1) data

N	Rejection Rate		Average Standard Error	RMSE	Standard Error of $\hat{\beta}$	Bias
	No effect	5 percent effect				
			A. CPS DATA			
50	0.058	0.822	0.0103	0.0106	0.0106	-0.002
20	0.064	0.41	0.016	0.016	0.016	0.019
10	0.0814	0.222	0.021	0.023	0.023	-0.06
6	0.1012	0.16	0.026	0.029	0.029	0.018
			B. HOMOGENOUS AR (1) DATA ($\rho = 0.8$)			
50	0.055	0.049	0.32	0.32	0.32	0.004
20	0.061	0.054	0.496	0.507	0.507	-0.005
10	0.0782	0.072	0.677	0.71	0.71	-0.004
6	0.0966	0.0906	0.823	0.929	0.929	-0.02

Notes*

1. In panel A, data from the Current Population survey has been used. The data is for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1980 to 2000 inclusive. The data has been aggregated as outlined in section 2.1 of the paper. No effect represents the type 1 error and 5% effect represents the power of the test. For the type 1 error (No effect) the data is aggregated, hence, the number of simulations is 5000, and for the 5% effect (power) the number of simulations is 500. This is because the log point is implemented at the micro level of the data making it computationally very expensive and time-consuming to run a higher number of simulations.
2. All simulations have been performed for the significance level of 5%
3. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.
4. For panel B, the data has been generated using a homogenous AR(1) structure. This analysis uses only the autocorrelation parameter, rho 0.8. In this panel, all the simulations are run 5000 times because the data is generated in an aggregated form.
5. For panel A, we randomly choose the required number of states for each simulation without replacement. For panel B we just generate the data for the required number of states in the first place.
6. All regressions also include state and year fixed effects in addition to the intervention variable.
7. The average standard error is calculated by taking the mean of the standard error associated with the intervention variable for all the simulations.
8. The standard error of $\hat{\beta}$ is the standard error associated with the distribution of coefficient of the intervention variable obtained from the simulations.
9. The average standard error, RMSE, standard error of $\hat{\beta}$, and bias are reported for the simulations, calculating the rejection rate of no effect.

Table 6: : Wild Cluster Bootstrapping for CPS data and homogenous AR(1) data

N	Rejection Rate		Average Standard Error	RMSE	Standard Error of $\hat{\beta}$	Bias
	No effect	5 percent effect				
			A. CPS DATA			
50	0.043	0.812	0.017	0.016	0.016	-0.0006
20	0.043	0.386	0.027	0.027	0.027	0.0004
10	0.05	0.196	0.0387	0.0388	0.0388	0.002
6	0.056	0.14	0.049	0.049	0.049	0.0001
			B. HOMOGENOUS AR (1) DATA ($\rho = 0.8$)			
50	0.055	0.053	0.518	0.517	0.517	-0.001
20	0.051	0.057	0.818	0.838	0.838	-0.002
10	0.064	0.062	1.37	1.152	1.151	0.05
6	0.061	0.064	1.48	1.509	1.509	0.012

Notes*

1. In panel A, data from the Current Population survey has been used. The data is for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1980 to 2000 inclusive. The data has been aggregated as outlined in section 2.1 of the paper. No effect represents the type 1 error and 5% effect represents the power of the test.
2. For wild cluster bootstrapping the number of simulations for no effect is 1000 because for each simulation, this method iterates 400 resamplings making it a very computationally expensive method. For power, the number of simulations remain 500.
3. All simulations have been performed for the significance level of 5%
4. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.
5. For panel B, the data has been generated using a homogenous AR(1) structure. This analysis uses only the autocorrelation parameter, rho 0.8. In this panel, all the simulations are run 1000 times because the data is generated in an aggregated form but bootstrapping itself is a very cumbersome method iterating 400 times for the resampling procedure in each simulation.
6. For panel A, we randomly choose the required number of states for each simulation without replacement. For panel B we just generate the data for the required number of states in the first place.
7. All regressions also include state and year fixed effects in addition to the intervention variable.
8. The average standard error is calculated by taking the mean of the standard error associated with the intervention variable for all the simulations.
9. The standard error of $\hat{\beta}$ is the standard error associated with the distribution of coefficient of the intervention variable obtained from the simulations.
10. The average standard error, RMSE, standard error of $\hat{\beta}$, and bias are reported for the simulations, calculating the rejection rate of no effect.

Table 7: : Feasible Generalized Least Squares for CPS data and homogenous AR(1) data

N	Rejection Rate		Average Standard Error	RMSE	Standard Error of $\hat{\beta}$	Bias
	No effect	5 percent effect				
			A. CPS DATA			
50	0.045	0.848	0.0171	0.0167	0.0165	0.002
20	0.047	0.406	0.027	0.026	0.026	0.003
10	0.051	0.244	0.038	0.038	0.037	0.006
6	0.056	0.18	0.049	0.050	0.049	0.008
			B. HOMOGENOUS AR (1) DATA ($\rho = 0.8$)			
50	0.058	0.064	0.2	0.207	0.207	0.002
20	0.068	0.072	0.316	0.333	0.333	0.005
10	0.0708	0.077	0.446	0.483	0.482	0.015
6	0.0988	0.1052	0.571	0.656	0.656	0.012

Notes*

1. In panel A, data from the Current Population survey has been used. The data is for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1980 to 2000 inclusive. The data has been aggregated as outlined in section 2.1 of the paper. No effect represents the type 1 error and 5% effect represents the power of the test. For the type 1 error (No effect) the data is aggregated, hence, the number of simulations is 5000, and for the 5% effect (power) the number of simulations is 500. This is because the log point is implemented at the micro level of the data making it computationally very expensive and time-consuming to run a higher number of simulations.
2. All simulations have been performed for the significance level of 5%
3. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.
4. For panel B, the data has been generated using a homogenous AR(1) structure. This analysis uses only the autocorrelation parameter, rho 0.8. In this panel, all the simulations are run 5000 times because the data is generated in an aggregated form.
5. For panel A, we randomly choose the required number of states for each simulation without replacement. For panel B we just generate the data for the required number of states in the first place.
6. Here the normal FGLS methd is used using the hansen(nobc) command in stata.
7. All regressions also include state and year fixed effects in addition to the intervention variable.
8. The average standard error is calculated by taking the mean of the standard error associated with the intervention variable for all the simulations.
9. The standard error of $\hat{\beta}$ is the standard error associated with the distribution of coefficient of the intervention variable obtained from the simulations.
10. The average standard error, RMSE, standard error of $\hat{\beta}$, and bias are reported for the simulations, calculating the rejection rate of no effect.

Table 8: : BC Feasible Generalized Least Squares for CPS data and homogenous AR(1) data

N	Rejection Rate		Average Standard Error	RMSE	Standard Error of $\hat{\beta}$	Bias
	No effect	5 percent effect				
			A. CPS DATA			
50	0.044	0.812	0.017	0.0165	0.0165	-0.0003
20	0.044	0.358	0.026	0.026	0.026	0.0008
10	0.051	0.21	0.038	0.037	0.037	0.001
6	0.05	0.15	0.049	0.0489	0.0489	0.0001
			B. HOMOGENOUS AR (1) DATA ($\rho = 0.8$)			
50	0.05	0.057	0.206	0.205	0.205	-0.007
20	0.048	0.05	0.307	0.302	0.302	0.007
10	0.062	0.066	0.46	0.47	0.47	0.02
6	0.072	0.077	0.060	0.064	0.064	0.02

Notes*

1. In panel A, data from the Current Population survey has been used. The data is for women between 25 and 50 in the fourth interview month of the Merged Outgoing Rotation Group for the years 1980 to 2000 inclusive. The data has been aggregated as outlined in section 2.1 of the paper. No effect represents the type 1 error and 5% effect represents the power of the test. For the type 1 error (No effect) the data is aggregated, hence, the number of simulations is 5000, and for the 5% effect (power) the number of simulations is 500. This is because the log point is implemented at the micro level of the data making it computationally very expensive and time-consuming to run a higher number of simulations.
2. All simulations have been performed for the significance level of 5%
3. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.
4. For panel B, the data has been generated using a homogenous AR(1) structure. This analysis uses only the autocorrelation parameter, rho 0.8. In this panel, all the simulations are run 2000 times because the data is generated in an aggregated form.
5. For panel A, we randomly choose the required number of states for each simulation without replacement. For panel B we just generate the data for the required number of states in the first place.
6. Here the normal FGLS methd is used using the hansen command in stata.
7. All regressions also include state and year fixed effects in addition to the intervention variable.
8. The average standard error is calculated by taking the mean of the standard error associated with the intervention variable for all the simulations.
9. The standard error of $\hat{\beta}$ is the standard error associated with the distribution of coefficient of the intervention variable obtained from the simulations.
10. The average standard error, RMSE, standard error of $\hat{\beta}$, and bias are reported for the simulations, calculating the rejection rate of no effect.

Table 9: Robustness Check: Misspecification of the error process

Method	Rejection Rate			
	Heterogeneous AR(1)		Homogeneous MA(1)	
	No Effect	5 percent	No Effect	5 percent
OLS (assuming iid)	0.37	0.388	0.153	0.135
CRSE $\mathcal{N}(0,1)$	0.051	0.05	0.063	0.064
CRSE t(g-1)	0.0424	0.0388	0.063	0.053
Residual Aggregation	0.092	0.0826	0.088	0.081
Wild Cluster Bootstrapping	0.055	0.057	0.054	0.047
FGLS	0.068	0.071	0.083	0.088
BC-FGLS	0.044	0.046	0.054	0.058

Notes*

1. The data has been generated separately for heterogenous AR 1 and homogenous MA 1 process and the associated type 1 error and power is reported in this table.
2. All simulations have been performed for the significance level of 5%
3. The rejection rate of No effect represents the type 1 error and that of 5% effect represents the power of the test.
4. All regressions also include state and year fixed effects in addition to the intervention variable.